LABORATOIRE D'INFORMATIQUE FONDAMENTALE D'ORLEANS

Université d'ORLÉANS

Towards Explainable Clustering: A Constrained Declarative based Approach

Mathieu GUILBERT, Christel VRAIN, Thi-Bich-Hanh DAO
firstname.name@univ-orleans.fr

Université d'Orléans - LIFO

anr

InvolvD

Summary

Summary

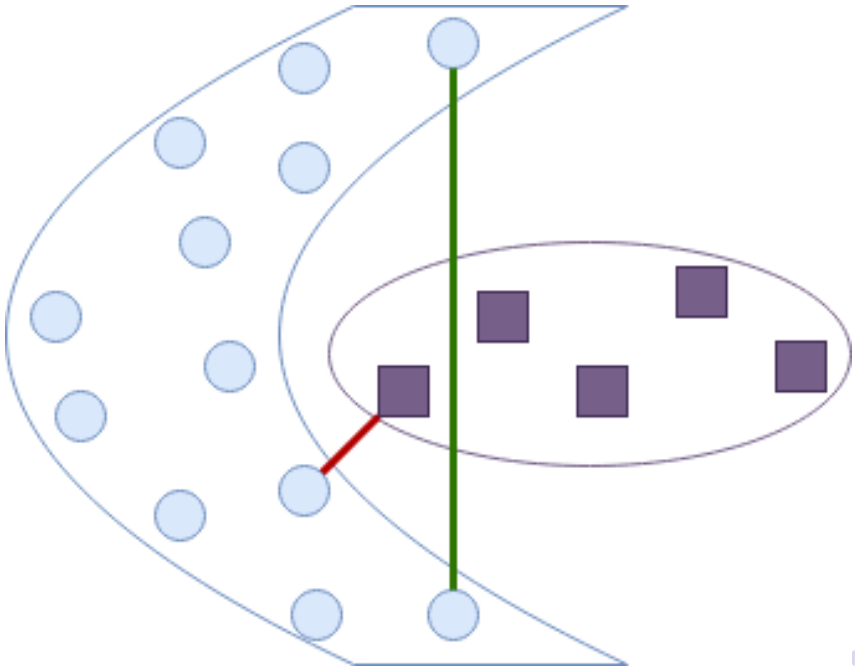
Constrained Clustering

Clustering

Find an underlying structure of a dataset by gathering data objects into groups called **clusters**, where points in the same group are more similar to each other than to those in other groups.

Incorporating prior knowledge in a clustering task in the form of constraints:

- instance level constraints: must-link (ML), cannot-link(CL)...
- cluster level constraints: diameter, separation,...



Interpretable clustering formulation

Data

$\mathcal{O}$: dataset composed of N instances.
$\mathbb{F}$: feature space.
$\mathbb{B}$: Boolean descriptor space.

$\mathbb{F}$ and $\mathbb{B}$ can be the same, overlapping or completely disjoint.

Goal

Find a clustering $\mathbb{P}$ of a dataset $\mathcal{O}$ where each cluster C_k is built according to the feature space and is associated to an explanation D_k composed of one or several covering and discriminative patterns $p \subseteq \mathbb{B}$.

2 options for the final clustering:

- Partition
- Overlapping clusters

Coverage

Pattern definition

A pattern $p \subset \mathbb{B}$ is a conjunction of descriptors, and an explanation D is a set of patterns.

Given a pattern p and an instance o, the following predicate defines whether p covers o:

$$cover(p, o) \stackrel{def}{=} \forall t \in p, \mathbb{B}_{o,t} = 1 \tag{1}$$

Pattern coverage definition

A pattern p *covers* a cluster C when it covers at least a percentage $\theta \in [0, 1]$ of its instances.

$$coverC(p, C, \theta) \stackrel{def}{=} \#\{o \in C \mid cover(p, o)\} \geq \theta . \#C \tag{2}$$

Discrimination

Discrimination definition

An explanation D_k of a cluster C_k is considered discriminative when *all* of its composing patterns are.

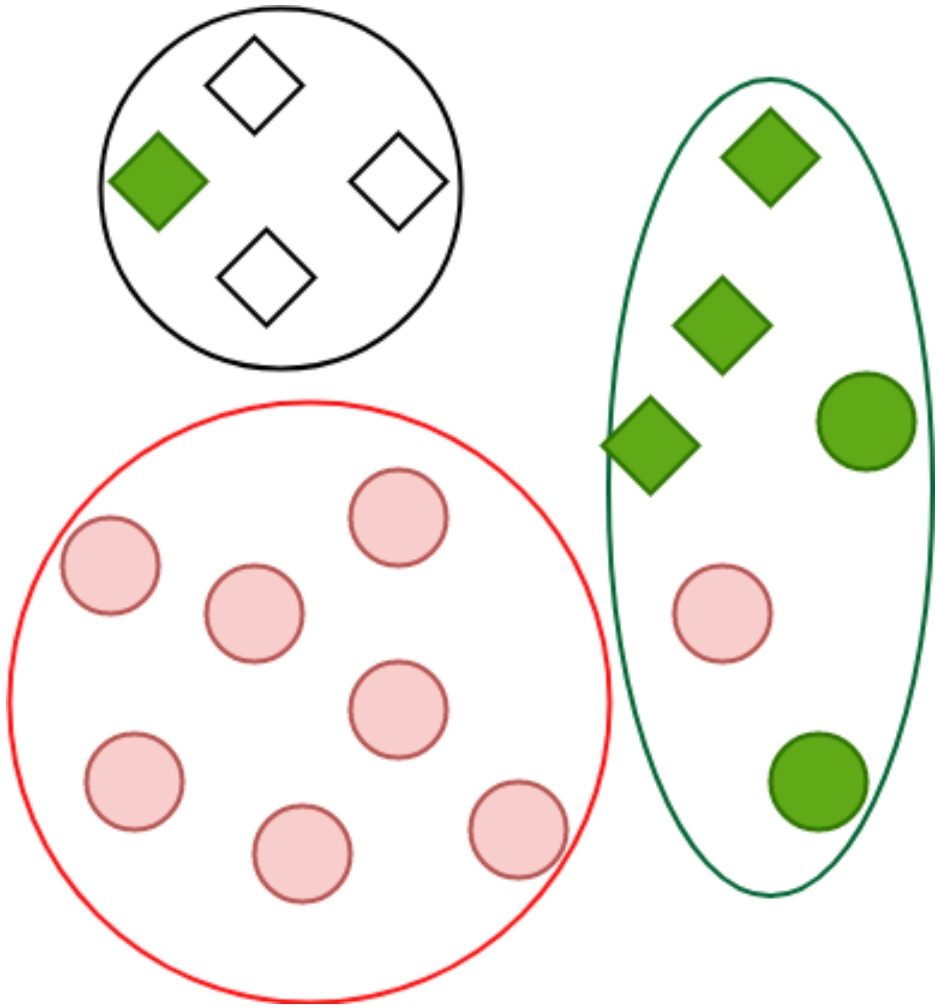
Cluster-wise pattern discrimination

- Each pattern p should not cover more than a certain amount ϕ of instances in each other cluster C' separately.

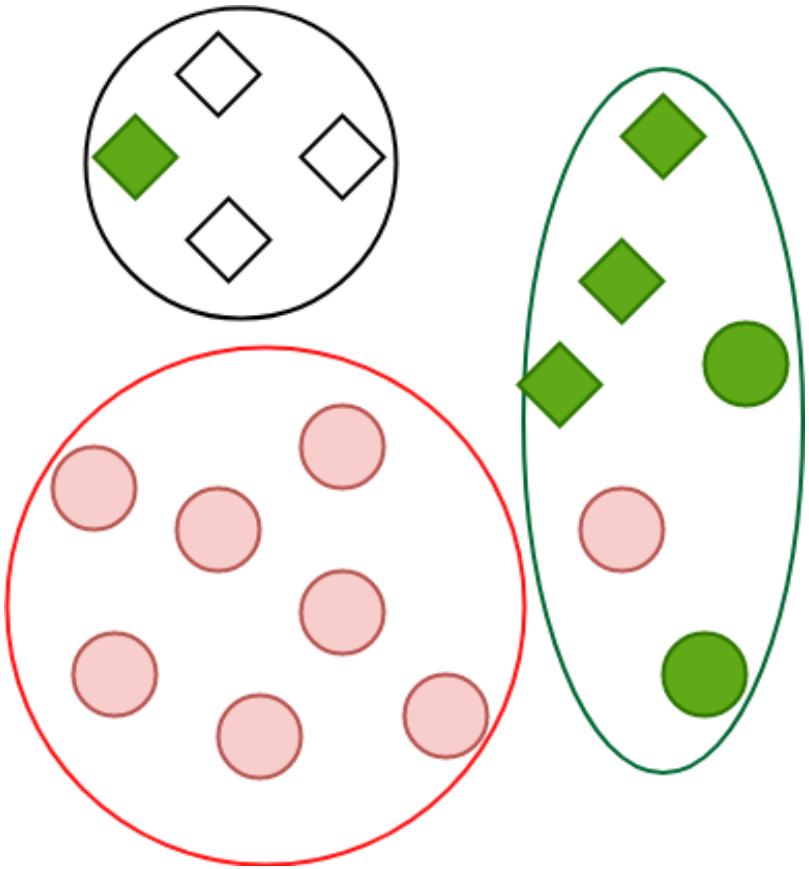
$$\forall C' \neq C,\ \#\{o \in C' \mid cover(p, o)\} < \phi.\#C' \tag{3}$$

Example

Given the following dataset of 17 points, with the colors and shapes as descriptors:



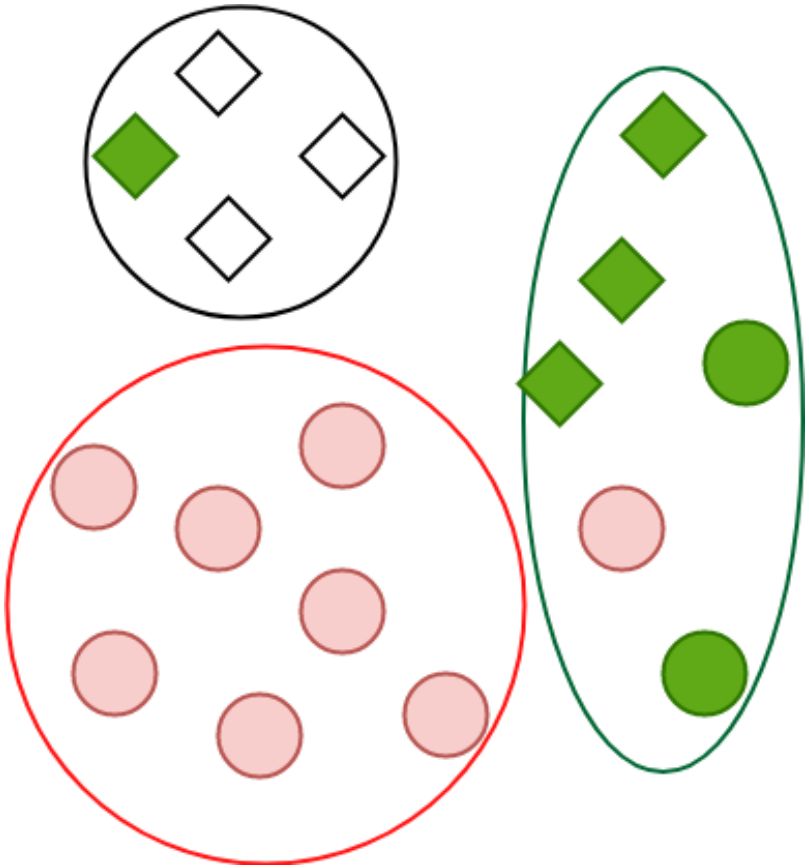
Example



With $\theta = 70$ and $\phi = 30$, explanations are the following:

- C0: *Red*
- C1: *Green*
- C2: *White*

Example

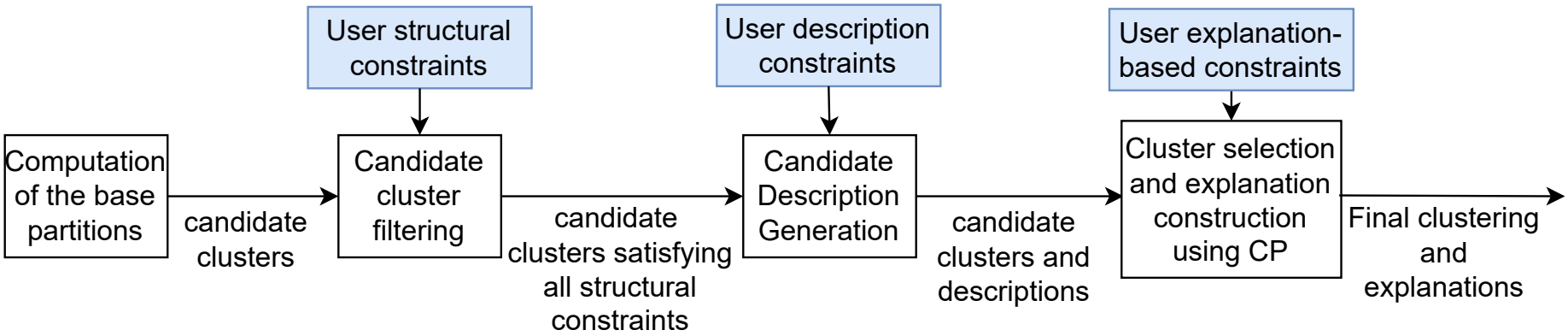


With $\theta = 70$ and $\phi = 30$ and conjunction as explanations:

- C0: *Red&Circle*
- C1: *Green*
- C2: *White&Square*

Summary

Overall approach



Building candidate clusters

Base partition generation options

- the same clustering algorithm with different parameters,
- different algorithms,
- different data representations,
- different subsets of the data,
- projection of the objects in different subspaces.

This step will generate many clusters, and some of them may not be worth of interest.
A filtering step is then performed to reduce the set of clusters.

Cluster Filtering

Selection options

- Keep the top % clusters in term of a criterion:
  - Within Cluster Sum of Square (WCSS)
  - diameter
  - ...
- Remove clusters not conforming to expert knowledge, in the form of user constraints.
  - Must-Link (ML) and Cannot-Link (CL)
  - Constraints on individual cluster (clusters sizes,...).

Candidate cluster descriptions

Associate to each cluster a description, i.e. set of patterns.

Pattern generation (Coverage)

2 options to generate patterns for each cluster C:

- Use single descriptors.
- LCM algorithm (Linear time Close itemset Miner), obtaining list of frequent closed patterns.

We keep only the clusters having at least one description pattern.

Constrained optimization problem

Inputs

- $V \in \mathbb{N}$: number of candidate clusters.
- $\mathbb{C} = \{C_1, ..., C_V\}$: with each $C_i \subseteq \mathcal{O}$ a candidate cluster C_i.
- $\mathbb{D} = \{D_1, ..., D_V\}$: with each $D_i \subseteq \mathcal{P}(\mathbb{B})$ a list of candidate patterns for cluster C_i.

Variables

- $\mathbb{P} \in \{0,1\}^V$: the final clustering, where $\mathbb{P}_i = 1$ means cluster C_i is selected and 0 otherwise.
- $\mathbb{Y} = \{\mathbb{Y}_1, ..., \mathbb{Y}_V\}$, each $\mathbb{Y}_i \in \{0,1\}^{|D_i|}$, where $\mathbb{Y}_{ij} = 1$ means the pattern $p_j \in D_i$ is used to describe C_i in the final clustering.

Clustering constraints

Number of clusters

$\mathbb{P}$ should have a number of clusters in a given range:

$$K_{min} \leq \sum_{c=1}^{V} \mathbb{P}_c \leq K_{max} \tag{4}$$

Number of attribution of instances

Let $nbClust(o)$ be the number of selected clusters that instance o is assigned to:

$$nbClust(o) = \sum_{c:o \in C_c} \mathbb{P}_c \tag{5}$$

We allow instances to be associated to multiple clusters, or no cluster at all:

$$\forall o \in \{1, .., V\}, \quad nbClustMin \leq nbClust(o) \leq nbClustMax \tag{6}$$

$$\sum_{o=1}^{N} (nbClust(o) \neq 1) \leq nbDiff1Max \tag{7}$$

Explanation constraints

Non-empty explanation

All selected clusters must have at least one of its candidate pattern selected to explain it.

$$\forall c \in \{1, .., V\}, \quad \mathbb{P}_c = 1 \implies \sum_{j=1}^{|\mathbb{Y}_c|} \mathbb{Y}_{cj} \geq 1 \tag{8}$$

Discrimination

For all pairs of clusters C_1 C_2, each pattern p describing C_1 should not cover more than ϕ percent of C_2's instances.

$$\#\{o \in C_2 \mid cover(p, o)\} \leq \phi \, \#C_2 \tag{9}$$

User constraints

Cluster selection user constraints

- *Must-Select*(c): forces the selection of a particular cluster c.

$$\mathbb{P}_c = 1 \tag{10}$$

- *Cannot-Select*(c_i, c_j): two particular clusters c_i and c_j cannot be both selected to create the clustering.

$$\mathbb{P}_{c_i} + \mathbb{P}_{c_j} < 2 \tag{11}$$

Objectives

- Maximizing the number of instances assigned to exactly one cluster:

$$f(\mathbb{P}) = \sum_{i=1}^{N} (nbClust(o) == 1) \tag{12}$$

- Minimizing the number of unassigned instances.

$$f'(\mathbb{P}) = \sum_{i=1}^{N} (nbClust(o) == 0) \tag{13}$$

- Maximizing or minimizing the sum of length of selected cluster explanations:

$$g(\mathbb{P}, \mathbb{Y}) = \sum_{c=1}^{V} (\mathbb{P}_c \times \sum \mathbb{Y}_c) \tag{14}$$

CP model

Inputs

- $N \in \mathbb{N}$: number of data instances
- $V \in \mathbb{N}$: number of candidate clusters
- $\mathbb{I} \in \{0,1\}^{V \times N}$: matrix where $\mathbb{I}_{ic} = 1$ if instance i is in cluster C_c, and 0 otherwise.
- $\mathbb{D} = \{D_1, ..., D_V\}$: with each $D_i \subseteq \mathcal{P}(\mathbb{B})$ a list of candidate patterns for cluster C_i.
- $\mathbb{W} \in \mathbb{N}^{V \times A}$: matrix where $\mathbb{W}_{cp}$ is the number of instances in cluster c that are covered by the pattern p, where A is the number of individual patterns.

Variables

- $\mathbb{P} \in \{0,1\}^V$
- $\mathbb{Y} = \{\mathbb{Y}_1, ..., \mathbb{Y}_V\}$
- $\mathbb{S} \in \mathbb{N}^N$: where $\mathbb{S}_i$ is the number of clusters in $\mathbb{P}$ containing instance i.

CP model - Discrimination constraints

- If a pattern j is selected in the final explanation of a selected cluster, then all other cluster for which j cover more then Θ elements cannot be selected.

$$\forall c = 1, \ldots, V, \forall j = 1, \ldots, |\mathbb{Y}_c|, \forall c' \neq c \text{ s.t.} \mathbb{W}_{c'j} \geq \Theta, \\ \mathbb{Y}_{jc} = 1 \implies \mathbb{P}_{c'} = 0 \quad (15)$$

- Every selected cluster as at least one descriptor.

$$\forall c = 1, \ldots, V, \ \mathbb{P}_c = 1 \iff \sum_{j=1}^{|\mathbb{Y}_c|} \mathbb{Y}_{cj} \geq 1 \quad (16)$$

- If a cluster is selected then all of its patterns that do not cover any other cluster are selected in its final explanation.

$$\forall c = 1, \ldots, V, \forall j \in \mathbb{D}_c, \mathbb{P}_c = 1 \wedge \bigwedge_{c' \neq c, \mathbb{W}_{c'j} > \Theta} \mathbb{P}_{c'} = 0 \implies \mathbb{Y}_{cj} = 1 \quad (17)$$

Summary

Evaluation metrics

- Pattern Coverage Rate (PCR), for a pattern p in the explanation D of a cluster C in clustering $\mathbb{P}$:

$$PCR(p, C) = \frac{\#\{o \in C : cover(p, o)\}}{\#C} \tag{18}$$

- Explanation Coverage (EC), measuring if instances are covered by at least one of their cluster's descriptive pattern:

$$EC(D, C) = \frac{\#\{o \in C \mid \exists p \in D \; cover(p, o)\}}{\#C} \tag{19}$$

- Inverse Pattern Contrastivity (IPC), evaluating the discrimination of a pattern instance-wise:

$$IPC(p, C_i) = \frac{1}{K-1} \sum_{C' \neq C_i \in \mathbb{P}} 1 - \frac{\#\{o \in C' : cover(p, o)\}}{|C'|} \tag{20}$$

Experimental setup

Code

The method has been implemented in Python 3, the CP model is built using CPMPY.

Patterns are generated with LCM (Linear time Close itemset Miner) algorithm, as implemented in scikit-mine Python library.

Unless specified otherwise, base partitions are generated with K-Means algorithm run twice with Euclidean distances and K ranging from 2 to 15.

For all the experiments, the explanation parameters are set as follows: Coverage(θ)=70%, Discrimination (ϕ)=30% and we accept overlapping between clusters.

Flags results

Table: Results on Flags with patterns of size 1.

C	**Explanation**	Size	PCR	EC	IPC
0	{Europe}	31	0.97	0.97	0.96
1	{Spanish}, {Catholic}	15	0.93	1.0	0.92
2	{NW}	34	1.0	1.0	0.86
3	{Oceania}	19	0.95	0.95	0.99
4	{Africa}	41	0.95	0.95	0.97
5	{Asia}	39	1.0	1.0	1.0

Flags results

Table: Results on Flags with LCM patterns.

C	**Explanation**	Size	PCR	EC	IPC
0	{Europe}, {Europe, NE}	29	0.98	1.0	0.98
1	{Catholic}	24	0.96	1.0	0.89
2	{small pop, NW}, {small area, small pop, NW}	27	0.98	1.0	0.96
3	{Oceania}	19	0.95	0.95	0.99
4	{Africa}	41	0.95	0.95	0.99
5	{Asia, NE}	39	1.0	1.0	1.0

Comparison

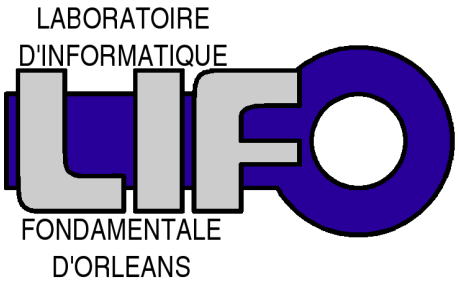
	Dataset	Iris	flags	AwA2
	K	3	6	2
Dao *et al.* 2018[1]	PCR	**1.0**	**1.0**	**1.0**
	EC	**1.0**	**1.0**	**1.0**
	IPC	0.77	0.62	0.38
K-Means	PCR	0.86	0.96	0.90
	EC	0.99	0.99	**1.0**
	IPC	0.73	0.70	0.39
K-Means (LCM)	PCR	0.83	0.86	0.87
	EC	0.99	**1.0**	**1.0**
	IPC	0.86	0.86	0.52
ECS	PCR	0.83	0.97	0.84
	EC	0.91	0.95	**1.0**
	IPC	0.88	**0.95**	**0.89**
ECS (LCM)	PCR	0.82	0.97	0.84
	EC	0.91	0.98	**1.0**
	IPC	**0.93**	**0.96**	**0.89**

[1] *Descriptive clustering: ILP and CP formulations with applications.* Dao, Thi-Bich-Hanh and Kuo, Chia-Tung and Ravi, SS and Vrain, Christel and Davidson, Ian. IJCAI 2018

Summary

ANR overview

ANR InvolvD: Interactive constraint elicitation for unsupervised and semi-supervised data mining



Website: https://involvd.greyc.fr/

Chemo-info data

- Data: 645 molecules
- First view: Inhibition percentage on 417 kinase proteins
- Second view: Presence/Absence of 12898 pharmacophores